De inzet van online peer assessment als formatief en summatief beoordelingsinstrument

van Boxel, P.; Reumer, C.G.; van Os, W

published in
Tijdschrift voor hoger onderwijs
2008

document version
Publisher's PDF, also known as Version of record

Patris van Boxel
Christoffel Reumer
Wim van Os
Jaap Boter

De inzet van online peer assessment als formatief en summatief beoordelings-instrument

Mw. drs. P.L.B. van Boxel (p.vanboxel@ond.vu.nl), drs. G.C. Reumer en dr. W. van Os zijn werkzaam bij het Onderwijscentrum VU van de Vrije Universiteit Amsterdam. Dr. J. Boter is docent aan de Faculteit der Economische Wetenschappen en Bedrijfskunde van de Vrije Universiteit Amsterdam.

Peer assessment, de procedure waarbij studenten elkaars werk van feedback voorzien of elkaar beoordelen, is een werk- en toetsvorm die momenteel sterk in de belangstelling staat. Vooral formatieve vormen van peer assessment worden in toenemende mate ingezet om de kwaliteit van het leren te bevorderen en studenten meer autonomie te geven in het leer- en beoordelingsproces. De voorbije jaren zijn specifieke softwaretoepassingen verschenen die de organisatie van peer-assessmentopdrachten ondersteunen, waardoor deze werkvorm ook bij grote studentgroepen praktisch haalbaar wordt.
In een exploratieve studie bij de Faculteit der Economische Wetenschappen en Bedrijfskunde (Vrije Universiteit Amsterdam) is ervaring opgedaan met de summatieve inzet en online ondersteuning van peer assessment bij een grote groep studenten. Uit de resultaten van de pilot blijkt dat studenten deze opzet als een leerzame manier beschouwen om leerstof te verwerken en feedback te krijgen over de eigen leerprestatie. Men is minder positief over het feit dat cijfers van medestudenten meetellen voor de eigen beoordeling. Maatregelen om de betrouwbaarheid van de studentoordelen te verhogen (zoals anonimiteit, scoringsrubrieken en cijferkalibratie), dragen in beperkte mate bij aan een hoger vertrouwen in het beoordelingsproces. De inzet van software blijkt essentieel om peer assessment op grote schaal te organiseren.

Inleiding

Leeropbrengsten van peer assessment

De verschuiving van kennisgericht naar competentiegericht opleiden gaat gepaard met de vraag om nieuwe beoordelingsvormen die verder gaan dan traditionele kennistoetsen. Dierick & Dochy (2001), Birenbaum (2003) en Segers (2004) beschrijven in dit verband de overgang van een testcultuur naar een assessmentcultuur, waarbij combinaties van toetsvormen worden ingezet om een breed scala aan kennis en vaardigheden door het leerproces heen te beoordelen. Kenmerkend voor de assessmentcultuur is onder meer dat beoordeling niet langer aan het einde van het leerproces plaatsvindt, maar onderdeel wordt van het leerproces zelf. Nieuwe toetsvormen hebben in de eerste plaats tot doel het leren te bevorderen en zijn dus gericht op kennisconstructie in

plaats van kennisreproductie. Daarbij wordt van studenten verwacht dat zij een actievere rol gaan spelen in het beoordelingsproces (Sluijsmans, 2002).
Peer assessment of peer review, de procedure waarbij studenten elkaars producten van commentaar voorzien of beoordelen volgens vooraf opgestelde beoordelingscriteria, is een werk- en toetsvorm die hierbij momenteel sterk in de belangstelling staat, zowel in onderzoek als in de onderwijspraktijk (Van den Berg, 2003; Dochy, Admiraal & Pilot, 2003; Segers, 2004). Studenten worden door verschillende leeractiviteiten in een peerassessmentopzet gestimuleerd om actiever met de leerinhoud bezig te zijn: ze krijgen beter inzicht in beoordelingscriteria, ze vergelijken hun eigen werk met dat van medestudenten, ze stellen sterke en zwakke punten vast in werk van anderen, verwerken dit in mondelinge of schriftelijke feedback en reageren op ontvangen feedback. Prins et al. (2005) beschouwen peer assessment in deze context ook als een specifieke vorm van samenwerkend leren.
Positieve leereffecten van peer assessment zijn onder meer de ontwikkeling van beoordelings- en communicatievaardigheden (Dochy, Admiraal & Pilot, 2003), een betere verwerking en begrip van de leerinhoud, verhoogd inzicht in de eigen leerprestatie en het stimuleren van reflectie op het eigen leergedrag (Dochy, Segers & Sluijsmans, 1999; Sluijsmans, 2002; Prins et al., 2005). Daarnaast worden efficiëntie en tijdbesparing als pragmatische redenen aangehaald om peer assessment in te voeren, vooral om grotere studentgroepen van just-in-time en gepersonaliseerde feedback te kunnen voorzien (Ballantyne, Hughes & Mylonas, 2002; Hamer, Ma & Kwong, 2005).

Betrouwbaarheid en validiteit van beoordelingen door peers
Waar in de literatuur vrij grote consensus bestaat over de positieve leereffecten van formatief peer assessment, met name die waarbij studenten elkaar kwalitatieve feedback geven tijdens het leerproces (Dochy, Segers & Sluijsmans, 1999; Sluijsmans, 2002; Topping, 2003), is er minder consensus waar het gaat om de betrouwbaarheid en validiteit van studentbeoordelingen die bestaan uit cijfers of scores. Cho, Schunn & Wilson (2006) concluderen op basis van hun reviewstudie dat er zowel voldoende theoretische grondslag is om de validiteit en betrouwbaarheid van peerbeoordelingen aan te nemen als af te wijzen. Een van de bevindingen uit hun empirische studie is dat meerdere peerbeoordelaars (vier tot zes) nodig zijn om een hoge mate van betrouwbaarheid te realiseren. Individuele studentbeoordelingen hebben een lagere betrouwbaarheid dan docentbeoordelingen. De keerzijde van het gebruik van meerdere beoordelaars is echter dat de kans groter is dat de feedback van elkaar afwijkt of inconsistent is. In de perceptie van studenten zijn verschillen tussen feedback van verschillende beoordelaars dan ook juist indicatief voor de lage betrouwbaarheid van peer feedback an sich. Cho, Schunn & Wilson (2006) concluderen dat student- en docentperspectieven op betrouwbaarheid (en validiteit) van een verschillende orde zijn, en te verklaren zijn door het feit dat de student alleen zicht heeft op de eigen beoordeling en de docent op die van de gehele studentenpopulatie.
Peer assessments zijn doorgaans minder betrouwbaar bij gebrek aan studentondersteuning in de vorm van training, checklists, voorbeelden en docentmonitoring (Topping, 2003). Sluijsmans (2002) wijst erop dat studenten begeleiding en training nodig hebben om de vaardigheden te verwerven die nodig zijn om de rol van beoordelaar op zich

te kunnen nemen. Meer ervaring met peer assessment blijkt dan te leiden tot meer vertrouwen in eigen kunnen en heeft een positieve invloed op het leereffect (Dochy, Admiraal & Pilot, 2003). Het is de vraag of hierdoor ook het vertrouwen in de beoordeling van peers kan worden versterkt, of dat ook andere maatregelen, zoals cijferweging, dat vertrouwen positief beïnvloeden. Segers (2004) pleit voor verder onderzoek rond peer assessment en de impact ervan op het leerproces, met name naar de mate waarin de peer assessor als 'fair' wordt beschouwd en naar de gepercipieerde validiteit van peer assessment.

Een valide beoordeling veronderstelt dat de gehanteerde beoordelingscriteria gepast zijn en juist gebruikt, dat wil zeggen: een juiste reflectie zijn van de criteria die experts gebruiken (Dierick, Dochy & Van de Watering, 2001). De validiteit van peer assessment wordt dan bepaald door de mate van overeenstemming tussen studentbeoordelingen en de beoordeling door een 'expert', in de meeste gevallen de docent van het vak. In Toppings review (1998) bleken 18 van de 25 studies waarin peers elkaar beoordelen een voldoende hoge validiteit te vertonen. Uit een meta-analyse van 48 onderzoeken naar peer assessment door Falchikov & Goldfinch (2000) bleek dat kennis en ownership van de beoordelingscriteria door de studenten leiden tot hogere consistentie tussen student- en docentbeoordelingen. Het gebruik van scoringsrubrieken (*rubrics*), een descriptief scoringsschema waarbij cijfers gerelateerd worden aan specifieke kwaliteitsindicatoren, kan daarbij een geschikte vorm zijn om studenten criteria te laten hanteren (Prins et al., 2005; Hamer, Ma & Kwong, 2005; Hamer, Kell & Spence, 2007).

Online ondersteuning van peer assessment

De voorbije jaren zijn specifieke softwaretoepassingen verschenen die de online organisatie van peer-assessmentopdrachten ondersteunen, zoals Turnitin, Espace, SWORD en Aropä. Bij grote groepen studenten is de automatische indiening en (random) toekenning van opdrachten aan beoordelaars een belangrijk logistiek voordeel voor de docent. Feedback en cijfers worden direct gegenereerd en voor de beoordeelde beschikbaar gesteld. De docent is in staat het volledige proces online te monitoren en kan in bijvoorbeeld hoorcolleges of werkgroepen inspelen op vaak voorkomende problemen bij het feedback geven (Tseng & Tsai, 2006).

Daarnaast kan de docent er in sommige softwareprogramma's voor kiezen het beoordelen anoniem door studenten te laten uitvoeren. Dit zorgt ervoor dat het beoordelingsproces objectiever kan verlopen: het voorkomt de verstorende impact van onder andere 'friendship marking' op de betrouwbaarheid van het beoordelingsproces (Hamer, Ma & Kwong, 2005; Tseng & Tsai, 2006) en heeft op die manier ook een positieve invloed op de perceptie van studenten omtrent de betrouwbaarheid van peer feedback.

Ten slotte laten online systemen ook toe dat peer assessment gemakkelijker ingezet kan worden bij afstandsonderwijs en computerondersteund samenwerkend leren. Hoewel verschillende vormen van computerondersteund peer assessment in de literatuur beschreven worden, is de impact van peer assessment op het leerproces tot dusver vooral gebaseerd op studies in face-to-face leersituaties. Daarnaast is kennis over het ontwerp van online peer-assessmentopdrachten, over procedures voor online samen-

werken en benodigde ondersteuning daarbij tot dusver schaars (Topping, 2003; Prins et al., 2005; Wen & Tsai, 2006). Datzelfde geldt voor onderzoek naar de betrouwbaarheid en validiteit van online peer assessment (Tseng & Tsai, 2006).

Onderzoeksvragen

Deze studie brengt de attitude in kaart van een grote groep studenten tegenover peer assessment in een online leeromgeving, waarbij de peer-assessmentopdrachten zowel een formatieve als summatieve focus hebben. De onderzoeksvragen richten zich op volgende drie aspecten van online peer assessment:

1. Wat is de attitude van studenten tegenover de inzet van peer assessment als werkvorm? Welke leereffecten ervaren studenten bij het uitvoeren van een reeks peer-assessmentopdrachten?
2. Welke attitude hebben studenten tegenover de inzet van peer assessment als toetsvorm?
   a. Hoeveel vertrouwen hebben studenten in hun medestudenten en in zichzelf als beoordelaar?
   b. Draagt weging van de cijfers bij aan een hoger vertrouwen in de *fairness* van summatief peer assessment en de gepercipieerde betrouwbaarheid?
3. Wat zijn de mogelijkheden en beperkingen van een software-tool om peer assessment te faciliteren bij grote studentgroepen?

Context

In 2006 is bij de Faculteit der Economische Wetenschappen en Bedrijfskunde van de Vrije Universiteit een pilot uitgevoerd met online peer-assessmentopdrachten bij een derdejaars bachelorvak Marketing (Consumer Behaviour). Aanvankelijk bestond dit vak uit hoorcolleges met werkgroepen en een afsluitend tentamen met open vragen. Door de sterke toename van het aantal studenten voor dit vak werden de werkgroepen afgeschaft. Aanleiding voor de pilot was de wens van de docent om de grote groep studenten op een meer actieve wijze met de leerstof te engageren en ze daarbij regelmatig van feedback te voorzien. Hierbij werd gekozen voor anoniem peer assessment als werken toetsvorm. Studenten gaven elkaar bij zes opdrachten wekelijks feedback aan de hand van scoringsrubrieken en een open vraag.

Deelnemers

Aan de cursus Consumer Behaviour namen 240 studenten deel met de volgende studieachtergrond:

- hbo-bachelors die een premasteropleiding volgen als voorbereiding op de masteropleiding Marketing (N = 165);
- bachelorstudenten economie (N = 25);
- bachelorstudenten bedrijfswetenschappen (N = 50).

Materialen

De studenten maakten over een periode van zes weken zes opdrachten. Alle opdrachten betroffen cases rond marketingaspecten van cultuur- en entertainmentproducten. Elke case refereerde aan één of twee hoofdstukken uit het cursushandboek en een actueel artikel. Afhankelijk van de case diende men extra literatuur te raadplegen of zelf te zoeken, kernbegrippen uit het handboek te identificeren en alternatieve marketingstrategieën of productvoorbeelden te bedenken. Elke case resulteerde in een Worddocument van ongeveer achthonderd woorden dat digitaal en anoniem ingeleverd werd.

Elke week diende men vervolgens de ingeleverde opdrachten van vijf medestudenten online en anoniem te beoordelen door middel van één open vraag en tien gesloten vragen.

De open vraag betrof een kwalitatieve evaluatie van de opdracht in zijn geheel. In minimaal vijftig woorden diende men commentaar te geven op de onderbouwing van de opdracht, originaliteit van de voorbeelden, sterke en zwakke punten en helderheid van het stuk.

De gesloten vragen dienden gescoord te worden op een schaal van één tot vijf. Vanaf week 3 introduceerde de docent scoringsrubrieken, waarbij elke score op de vijfpuntsschaal correspondeerde met een bepaalde kwaliteitsomschrijving door de docent. Nadat de deadline voor het leveren van commentaar was verstreken, konden studenten hun cijfers en feedback online raadplegen.

Cijferberekening en wegingsprocedure

Studenten werden geïnformeerd dat 40% van hun eindcijfer was gebaseerd op cijfers van medestudenten ontvangen tijdens de caseopdrachten; de resterende 60% van het eindcijfer was gebaseerd op een multiplechoicetentamen.

Het ontvangen cijfer van (vijf) medestudenten op een caseopdracht was een gewogen gemiddelde cijfer. De weging gebeurde door aan elke beoordelaar een individuele wegingsfactor toe te kennen. Hiervoor werd de beoordeling die iedere beoordelaar maakte, vergeleken met de beoordelingen van de andere beoordelaars die hetzelfde werkstuk hadden beoordeeld. Als de beoordeling overeenkwam met (het gemiddelde van) de andere beoordelingen, kreeg die beoordelaar een hoog gewicht en telde zijn beoordeling zwaarder mee in de gewogen beoordeling van het werkstuk dan de beoordeling van iemand die minder overeenkwam met de andere beoordelingen (meer afweek van de gemiddelde beoordeling). Deze wegingsprocedure is gebaseerd op Hamer, Ma & Kwong (2005) en is met name bedoeld om *rogue* (niet-serieuze) beoordelaars te identificeren en hun cijfer minder te laten wegen dan dat van serieuze beoordelaars. Volgens Hamer et al. is het namelijk aannemelijk dat elke groep beoordelaars enkele leden kent die willekeurige scores geven, en dat dit in het algemeen degenen zullen zijn die de meest afwijkende cijfers uitdelen. De overweging voor dat laatste is weer dat de gemiddelde score in een verdeling uiteindelijk de beste schatter is van de ware score.

De cijfercorrectie resulteerde gemiddeld in ongeveer 0,3 punt verschil met het ongewogen gemiddelde van de cijfers tussen één en vijf. De gemiddelde wegingsfactor die voor elke beoordelaar werd vastgesteld, werd nog eens verrekend in zijn of haar ontvangen cijfer voor een opdracht. In de discussie later in dit artikel komen we terug op

de vraag in hoeverre de aannames van Hamer, Ma & Kwong juist zijn, met daaraan gekoppeld de vraag of de weging ook van invloed is op de hoogte van de gegeven beoordelingen.
Studenten die niet tevreden waren met hun beoordeling, werden bij elke opdracht in de mogelijkheid gesteld om een herbeoordeling door de docent te krijgen. Wekelijks hebben vijf of minder studenten van deze mogelijkheid gebruikgemaakt.

Onderzoeksinstrument
Na afloop van het tentamen is een schriftelijke vragenlijst afgenomen bij studenten. Het doel van dit instrument was:
1. Inzicht krijgen in de attitude van studenten tegenover de specifieke inzet van peer assessment als werkvorm binnen de cursus Consumer Behaviour (onderzoeksvraag 1), met name door statements over verschillende leereffecten van het uitvoeren van peer-reviewopdrachten voor te leggen.
2. Inzicht krijgen in de attitude van studenten tegenover de specifieke inzet van peer assessment als toetsvorm (onderzoeksvraag 2). Daarbij zijn vragen voorgelegd over het vertrouwen dat men heeft in peers als assessors en over de inzet van cijferweging als manier om het vertrouwen in de totstandkoming van het eindcijfer te vergroten.
3. Inzicht krijgen in de mogelijkheden en beperkingen van de inzet van een softwaretool ter ondersteuning van peer-assessmentopdrachten (onderzoeksvraag 3).

De vragenlijst bestond uit 24 gesloten vragen (op een vijfpuntsschaal), drie meerkeuzevragen en één open vraag:
- De 24 Likert-schaalvragen bevatten statements over de leereffecten van peer assessment, de procedure voor cijferweging, de eigen beoordelingsvaardigheid en die van medestudenten en de gebruikte software. Appendix A bevat de letterlijke tekst van deze vragen.
- De drie meerkeuzevragen waren bedoeld om de studieachtergrond (hbo-premasters, bachelors economie of bachelors bedrijfswetenschappen), de voorkeur voor de huidige of oorspronkelijke cursusopzet en de gewenste onderwijsvorm (de huidige opzet of de vroegere) vast te stellen.
- Bij de open vraag konden studenten omschrijven welke waarde zij toekennen aan het gebruik van peer assessment voor hun eigen ontwikkeling.

De vragenlijst is na het tentamen ingevuld door 164 studenten (68,3% van 240). Het ging daarbij om 92 hbo-studenten (55,8% van 165), 24 bachelors economie (96,0% van 25) en 26 bachelors bedrijfswetenschappen (52,0% van 50). Van in totaal 22 studenten is de studieachtergrond niet bekend (zij hebben de vraag niet beantwoord). Het is duidelijk dat de responspercentages sterk verschillen, afhankelijk van de studieachtergrond. In de paragraaf ‘Effect van studieachtergrond’ wordt dit verder besproken.

Data-analyse
Van de 24 vijfpuntsvragen werden het gemiddelde en de standaarddeviatie berekend, mede uitgesplitst naar studieachtergrond. Ten behoeve van de bespreking in de paragraaf ‘Resultaten’ zijn twee groepen van vragen geformeerd, te weten:

- de opvattingen van studenten over de waarde van peer assessment als werkvorm (zeven vragen); de betrouwbaarheid (Cronbach's α) van deze vraaggroep bedraagt 0,72;
- de opvattingen van studenten over peer assessment als beoordelingsinstrument (eveneens zeven vragen), met een betrouwbaarheid van 0,67.

Mede gezien het geringe aantal items kan in beide gevallen gesproken worden van een aanvaardbare betrouwbaarheid (Nunnally, 1978: 245). Welke hoogte acceptabel is, hangt overigens ook af van de vraag waarvoor een en ander is bedoeld, en die grenzen liggen vanzelfsprekend hoger in selectieve situaties (personeelsbeoordelingen, studietoetsen) dan in de onderhavige, waarin het er vooral om gaat in hoeverre de desbetreffende items 'echt' bij elkaar horen.
Om inzicht te krijgen in de spreiding en ontwikkeling van de beoordelingen, zijn de beoordelingen van week 2 tot 5 geanalyseerd, alsmede de beoordelaarscoëfficiënten. De gemiddelden en standaarddeviaties zijn berekend over de hele groep en binnen beoordelingsgroepen (dat wil zeggen beoordelaars van eenzelfde werkstuk). De gemiddelden en standaardafwijkingen binnen beoordelingsgroepen geven inzicht in het beoordelingsgedrag van de studenten (ontwikkeling van de aan peers toegekende cijfers in de loop van de cursus) en de mate van overeenstemming tussen de beoordelaars van eenzelfde opdracht.

Resultaten

Tabel 1 bevat de uitkomsten van de vragen over peer assessment als werkvorm. De vragen in kwestie zijn in de tabel van boven naar beneden gerangschikt in die zin dat de meest positief beantwoorde vraag bovenaan staat en de minst positieve onderaan. De vraagnummers verwijzen naar de nummering in appendix A. Bij elke vraag loopt de codering van - - (code 1, zeer oneens) tot ++ (code 5, zeer eens).

Tabel 1 Peer review als werkwijze en leereffect (α = .72)

	- -	-	+/-	+	++	Gem	Sd	N
17 Commentaar medestudenten gelezen	3,8%	15,7%	14,5%	33,3%	32,7%	3,75	1,18	159
5 Peer review goede methode	5,6%	14,4%	20,0%	51,9%	8,1%	3,43	1,02	160
22 Beoordelen medestudenten leuke werkwijze	5,6%	11,9%	35,6%	41,3%	5,6%	3,29	0,95	160
18 Vergelijking antwoorden leerzaam	3,1%	18,1%	35,0%	38,8%	5,0%	3,24	0,92	160
19 Formuleren commentaar leerzaam	3,8%	23,3%	39,6%	30,2%	3,1%	3,06	0,90	159
20 Lezen commentaar medestudenten leerzaam	7,6%	25,5%	38,2%	26,8%	1,9%	2,90	0,95	157
16 Commentaar medestudenten goed onderbouwd	8,8%	32,7%	51,6%	6,9%	0,0%	2,57	0,75	159

Studentattitude tegenover peer assessment als werkvorm, en het ervaren leereffect

Ongeveer twee derde deel heeft (vrijwel) wekelijks kennisgenomen van het commentaar van medestudenten op de gemaakte opdrachten (item 17). Minder dan 4% heeft dat naar eigen zeggen zelden of nooit gedaan, en de rest (ongeveer 20%) kennelijk af en toe. Dat betekent niet dat men zo positief oordeelt over dat commentaar (vraag 20). Minder dan 30% vond het (zeer) leerzaam, en voor de overige 70% gold dat (veel) minder. Dat zal vermoedelijk samenhangen met het feit dat slechts 7% het commentaar tamelijk goed onderbouwd vond, ruim de helft is hierover wisselend gestemd en de overige 40% is in dit verband in meerdere of mindere mate negatief (item 16). Dat het overall oordeel over peer assessment (items 5 en 22) toch redelijk gunstig uitvalt, zal dan ook meer te maken hebben met het positieve leereffect dat studenten ervaren bij het geven van feedback en het vergelijken van het eigen werk met dat van anderen (items 18 en 19), dan met het ontvangen van feedback en de gepercipieerde intrinsieke betekenis van het commentaar van medestudenten.
Studenten konden ook via een open vraag aangeven welke waarde zij toekennen aan peer assessment voor hun eigen ontwikkeling. Deze open vraag werd beantwoord door 79% van de respondenten. De antwoorden werden ondergebracht in zes antwoordcategorieën (tabel 2).

Tabel 2 Waarde van peer review

Wat is de waarde van peer review voor je eigen ontwikkeling?	**N**	**%**
Je leert van antwoorden van anderen	27	21
Actiever, dieper, langer met de leerstof bezig, betere beheersing van stof	20	15
Anderen beoordelen is leerzaam	20	15
Het is interessant/leerzaam te weten wat anderen van je werk vinden	18	14
Andere redenen: leuke werkvorm, schrijfstijl ontwikkelen, betere tentamenvoorbereiding, toepassen theorie op praktijk, …	13	10
Ik vind het niet waardevol	32	25
	130	100

Van de studenten die de open vraag beantwoordden, kende 75% een positieve waarde toe aan de inzet van peer assessment voor hun eigen ontwikkeling. De antwoorden op deze open vraag vertonen consistentie met de resultaten van de gesloten vragen over de leerzaamheid van diverse aspecten van het beoordelingsproces. Bij de peer-assessmentopzet in deze pilot was het meest leerzame onderdeel het bekijken van uitwerkingen van anderen (en het vergelijken met de eigen uitwerking). Ook het intensiever (beter, dieper en langer) met de leerstof bezig zijn en het (leren) beoordelen van anderen, werd door veel studenten als waardevol omschreven.
Het krijgen van de feedback van anderen werd blijkens tabel 1 door minder studenten als leerzaam beschouwd. Studenten die deze opzet minder of niet waardevol vonden

(25%, zie tabel 2), haalden vooral redenen aan zoals het krijgen van een 'onjuiste' beoordeling, 'onzincommentaar' van medestudenten en een gebrek aan vertrouwen in de beoordelingsmethode.

Voorkeuren voor cursusopzet en studiemethode
De twee meerkeuzevragen gaven meer inzicht in de voorkeuren van de studenten rond gebruikte werk- en toetsvormen in het vak, en voorkeuren voor een meer of minder 'activerende' werkmethode. Bij de vraag 'Verkies je de huidige opzet van het vak (hoorcolleges met zes opdrachten, dertig peer reviews en een multiplechoicetentamen) of de eerdere opzet (hoorcolleges, geen opdrachten en een tentamen met open vragen)?' kiest 87% van de respondenten voor de huidige opzet. Bij de vraag of het 'regelmatig met de stof bezig zijn gedurende de cursus' dan wel het 'kort en intensief leren van de stof voorafgaand aan het tentamen' de voorkeur krijgt, kiest 83% voor het regelmatig met de stof bezig zijn.

Studentattitude tegenover de summatieve inzet van peer assessment

Slechts 13% van de respondenten staat positief tot zeer positief tegenover het statement 'Ik vind peer review een goede methode om mijn cijfer vast te stellen' (tabel 3, item 6). Dit wijkt sterk af van de positieve attitude van 60% van de respondenten tegenover peer assessment als werkvorm (tabel 1, item 5). Men zou dit ook zo kunnen formuleren: peer assessment als werkvorm en manier om (mede) het cijfer te bepalen is volgens de student wel geschikt voor hemzelf, maar niet, of in elk geval minder, voor zijn medestudenten.

Tabel 3 Peer review versus cijfers geven en ontvangen (α = .67)

	- -	-	+/-	+	++	Gem	Sd	N
14 Zeker van juist cijfer aan medestudenten	1,9%	4,4%	20,6%	61,9%	11,3%	3,76	0,78	160
9 Beoordelingscriteria duidelijk	1,8%	6,7%	20,2%	60,7%	10,4%	3,71	0,81	163
7 Verschillende weging beoordelaars verstandig	3,7%	6,1%	31,3%	47,2%	11,7%	3,57	0,91	163
4 Totstandkoming eigen cijfer duidelijk	5,7%	18,2%	30,2%	39,0%	6,9%	3,23	1,01	159
15 Scores van medestudenten adequaat	8,8%	24,4%	43,8%	22,5%	0,6%	2,82	0,90	160
8 Vertrouwen in juistheid van mijn cijfer	13,7%	28,6%	34,2%	22,4%	1,2%	2,69	1,01	161
6 Peer review goede methode t.a.v. mijn cijfer	18,6%	31,1%	37,3%	13,0%	0,0%	2,45	0,94	161

Vertrouwen in eigen beoordelingsvaardigheid en die van medestudenten
Studenten hebben een groot vertrouwen in hun eigen beoordelingsvaardigheid: 73% is het eens tot zeer eens met het statement 'Ik ben er zeker van dat ik een juiste score heb toegekend aan mijn medestudenten' (tabel 3, item 14). In schril contrast hiermee staat de tevredenheid over de cijfers die men van medestudenten kreeg: slechts 23% is het ermee eens dat dit adequate cijfers zijn (tabel 3, item 15). In dezelfde lijn ligt de beantwoording van de vraag over het 'vertrouwen in de juistheid van het verkregen cijfer' (tabel 3, item 8), en de meer algemene vraag of peer assessment wel een goede methode is om het cijfer vast te stellen (item 6).

Cijferweging
Gezien het voorgaande ligt het in de rede dat studenten voorstander zijn van de toegepaste weging waardoor extreme, 'niet-serieuze' waarderingen niet meetellen: 59% van de respondenten is het ermee eens (tabel 3, item 7). Hierdoor heeft 23% ook meer vertrouwen in de totstandkoming van het cijfer (tabel 3, item 8), maar voor ruim 42% is dit – zoals hierboven al bleek – toch niet of slechts beperkt het geval.
Niet in de tabel opgenomen is de vraag of studenten gematigder cijfers zijn gaan geven omdat er een weging werd uitgevoerd (wat indirect invloed had op hun eindcijfer aangezien daarin ook de beoordelaarskwaliteit meespeelde). Hierop wordt zeer uiteenlopend gereageerd: 47,5% vindt van wel, 27% denkt van niet en een kwart scoort neutraal op deze vraag. Zoals in de paragraaf 'Cijferberekening en wegingsprocedure' al is aangegeven, komen we in de discussie op dit punt terug.
De cijfers die studenten aan elkaar gaven, zijn geanalyseerd voor week 2 tot en met 5. Daarbij is nagegaan of studenten gaandeweg hoger of lager beoordelen en hoe de overeenstemming tussen de beoordelaars van eenzelfde werkstuk zich door de opdrachtenreeks heen ontwikkelt. In tabel 4 wordt het globaal gemiddelde cijfer per caseopdracht per week weergegeven. Hieruit blijkt dat het gemiddelde cijfer afneemt met 0,19 punt tussen week 2 en 3 (van 3,49 naar 3,30), maar daarna vrijwel constant blijft.

Tabel 4 Gemiddelde cijfer van de peerbeoordelingen van week 2 tot en met 5 (Anova; F = 46,097; df = 3; p < 0,000)

	Cijfer		
Week	**Gemiddelde**	**Sd**	**N**
2	3,49	0,54	943
3	3,30	0,50	1034
4	3,27	0,47	1041
5	3,28	0,39	1054

De variantieanalyse geeft aan dat het overall verschil tussen de weken 2 tot en met 5 significant is ($p < 0{,}000$), maar de post-hoc Bonferroni-toets laat zien dat week 2 wel verschilt ten opzichte van week 3, 4 en 5, maar dat de drie laatstgenoemde weken onderling niet significant verschillen.

Daarnaast neemt de overeenstemming tussen de studentbeoordelingen toe (van 0,198 naar 0,100 uitgedrukt als het kwadraatverschil binnen beoordelingsgroepen): de afwijking tussen de cijfers die beoordelaars aan eenzelfde opdracht geven, wordt namelijk minder naarmate de opdrachtenreeks vordert (tabel 5).

Tabel 5 Gemiddeld gekwadrateerd verschil tussen het cijfer dat een beoordelaar geeft en het gewogen gemiddelde van de cijfers binnen een beoordelingsgroep (Anova; F = 23,467; df = 3; p < 0,000)

Week	Gemiddelde kwadraatverschil	Sd	N
2	0,198	0,39	943
3	0,156	0,28	1034
4	0,124	0,23	1041
5	0,100	0,18	1054

Ook hiervoor geldt een overall significant verschil tussen de diverse weken ($p < 0{,}000$), dat gezien de post-hoc toets ook voor de weken onderling gehandhaafd blijft, tot en met week 4: week 5 verschilt niet meer significant van de voorgaande week.

Effect van studieachtergrond

De studieachtergrond van de studenten houdt verband met de waardering van online formatief en summatief peer assessment (tabel 6). Studenten met een hbo-achtergrond zijn vrijwel steeds positiever over het gebruik van peer review als werkvorm en als beoordelingsinstrument dan bachelorstudenten, met name die van economie. In tabel 6 geven we een overzicht van de vragen waarin het verschil significant is.
Het is niet zonder meer duidelijk waaraan een en ander moet worden toegeschreven. In beginsel is het mogelijk dat de afwijkende respons van de verschillende groepen hierbij een rol heeft gespeeld. Zoals in de paragraaf 'Onderzoeksinstrument' is aangegeven, was de respons van hbo-studenten en die van studenten bedrijfswetenschappen 50 tot 55%, die van studenten economie bijna 100%. Het is denkbaar dat naar verhouding de groep hbo-studenten en bachelors bedrijfswetenschappen uit meer succesvolle studenten bestond, terwijl dat minder het geval was bij de bachelors economie. Wellicht zijn succesvolle studenten positiever ten opzichte van peer assessment omdat zij hogere cijfers krijgen. Daar staat echter tegenover dat in de data nergens sterke correlaties zijn aangetroffen tussen de waardering voor peer review enerzijds en de hoogte van het verkregen cijfer voor de cases anderzijds. Bovendien blijft ook na correctie voor het ontvangen cijfer via een uitgevoerde covariantieanalyse het verschil tussen studenten met deze of gene studieachtergrond bestaan. In de discussie komen we hier kort op terug.

Tabel 6 Gemiddelden op Likert-schaalitems uitgesplitst naar studieachtergrond (Anova; * = p < 0,05; ** = p < 0,01)

Item	HboO	BAeco	BAbws	Totaal	P
5 Ik vind peer review een goede methode om met de leerstof bezig te zijn	3,63	3,00	3,38	3,43	.020*
6 Ik vind peer review een goede methode om mijn cijfer vast te stellen	2,65	2,23	2,23	2,45	.041*
7 Ik ben het ermee eens dat beoordelaars verschillend gewogen worden in de beoordeling van mijn case	3,73	3,04	3,50	3,57	.006**
8 Ik heb er vertrouwen in dat doordat mijn cijfer voor een case verschillend wordt gewogen per beoordelaar, het juiste cijfer tot stand komt	2,92	2,35	2,38	2,69	.008**
18 Ik heb veel geleerd van het vergelijken van mijn antwoorden op de cases met de antwoorden van mijn medestudenten	3,42	2,96	2,96	3,24	.021*
20 Ik heb veel geleerd van het lezen van commentaar van medestudenten op mijn eigen opdrachten	3,07	2,87	2,40	2,90	.007**
22 Ik vond het leuk om door middel van het beoordelen van cases van medestudenten met de leerstof bezig te zijn	3,49	3,14	3,00	3,29	.032*

Online ondersteuning van peer-assessmentopdrachten

Het gebruik van de software heeft geregeld voor problemen gezorgd. Zo zijn na afloop van de eerste opdracht alle antwoorden op de open vragen verdwenen. Dit heeft er mede toe geleid dat de eerste opdracht niet heeft meegeteld en beschouwd werd als een oefenopdracht. De software voorzag niet in mogelijkheden tot automatische cijferweging, zodat dit handmatig door de docent diende te worden uitgevoerd, wat zeer tijdsintensief was.
Problemen met de bereikbaarheid en stabiliteit van de software leken niet van invloed op de evaluatie van het gebruiksgemak of de algemene waardering van een online peer-assessmentopzet. Van de respondenten is 93% positief over de gebruiksvriendelijkheid van de peer-reviewsoftware, en eerder bleek al dat de waardering voor peer review als werkvorm (tabel 1, item 5, en tabel 2) over het algemeen positief was.

Discussie

De inzet van peer assessment als werkvorm
De eerste onderzoeksvraag was gericht op het in kaart brengen van attitudes van studenten tegenover de inzet van peer assessment als een werkvorm en de leereffecten die studenten daarbij percipiëren. Bijna twee derde van de respondenten is positief over de inzet van peer assessment als werkvorm. Dat een groot deel van deze studenten

positief staat tegenover een activerende werkvorm als peer assessment, sluit aan bij een overheersende voorkeur voor een actieve studiemethode en de huidige cursusopzet met veel opdrachten en peer reviews. Het meest waardevolle leereffect bij deze peer-assessmentopzet bleek het lezen van uitwerkingen van anderen en het (impliciet) vergelijken daarvan met het eigen werk. Dit gepercipieerde leereffect is ook beschreven door onder andere Ballantyne, Hughes & Mylonas (2002) en Wen & Tsai (2006).
Veel studenten waren verder van mening dat de reviewopdrachten hadden geleid tot een betere beheersing van de leerstof. Segers & Dochy (2001) stelden eerder dit gepercipieerde leereffect vast van het ontwikkelen van zogenaamd 'dieper leergedrag'.
Het formuleren van commentaar op het werk van anderen werd als leerzamer gezien dan het ontvangen van commentaar van anderen. De kwaliteit van de feedback van medestudenten vond men over het algemeen vrij laag. De redenen hiervoor zijn waarschijnlijk het directe gevolg van de wijze waarop deze opdracht is ontworpen: de ontvangen feedback (op basis van één open vraag) diende men niet actief te verwerken in een latere versie van dezelfde opdracht. In het gunstigste geval lazen studenten de ontvangen feedback na, maar dit was geen verplichting. De kwaliteit van de feedback werd ook niet beoordeeld, noch door de gereviewde student, noch door de docent. De extrinsieke motivatie om kwalitatief goede feedback te ontwikkelen ontbrak dus in deze peer-assessmentopzet, waardoor ook een deel van het mogelijke leereffect van peer assessment in deze opzet niet werd gerealiseerd. Ballantyne, Hughes & Mylonas (2002) pleiten ervoor om een redelijk deel van het cijfer (10 tot 15%) te reserveren voor de uitvoering én kwaliteit van de peer assessment om ervoor te zorgen dat studenten de taak serieus uitvoeren. Ook ontvingen studenten geen beoordelingstraining bij aanvang van het vak, volgens Sluijsmans (2002) toch een belangrijke voorwaarde om toereikende beoordelingsvaardigheden te kunnen ontwikkelen.

De summatieve inzet van peer assessment
De tweede onderzoeksvraag was gericht op het in kaart brengen van attitudes van studenten tegenover de inzet van peer assessment voor summatieve doeleinden. Daarbij is aan studenten gevraagd hoeveel vertrouwen ze in hun medestudenten en in zichzelf hebben als beoordelaar, en of de toegepaste cijferweging op de studentoordelen de (gepercipieerde) betrouwbaarheid positief beïnvloedt. Uit de resultaten blijkt dat studenten negatiever denken over de inzet van peer assessment als toetsinstrument dan als werkvorm of leeractiviteit.
Eén reden voor de sceptische attitude van sommige studenten tegenover beoordelingen door peers is het gebrek aan vertrouwen in de beoordelingscapaciteit van medestudenten. Dit beperkte vertrouwen werd mogelijk versterkt doordat een relatief groot deel van het eindcijfer (40%) gebaseerd was op studentbeoordelingen. Wen & Tsai (2006) beschrijven in dit verband de attitude van studenten tegenover de verhouding van de peer-assessmentbeoordeling tot de totale eindbeoordeling. Studenten staan over het algemeen positief tegenover een beperkte opname van studentbeoordelingen in de eindbeoordeling, en negatief tegenover studentbeoordelingen wanneer die voor de helft of meer meetellen voor het eindcijfer. In deze studie is studenten niet gevraagd naar hun mening hierover, maar duidelijk is dat de verhouding tussen peer- en docent-

beoordelingen voor het genereren van een eindcijfer voorzichtig dient te worden afgewogen bij de opzet van peer-assessmentopdrachten.
Het lage vertrouwen in beoordelingen van peers vertoont enigszins een discrepantie met het hoge vertrouwen dat studenten hebben in zichzelf als beoordelaar. Er zijn weinig studies bekend waarin de factor 'vertrouwen in eigen beoordelingsvaardigheid' is onderzocht. Dit pleit voor meer onderzoek naar de redenen van deze discrepantie in studentpercepties.
Ondanks het invoeren van maatregelen om het beoordelingsproces objectiever en betrouwbaarder te maken (zoals de cijferweging, het anonimiseren van de beoordelingen, de inzet van meerdere beoordelaars per werkstuk én het gebruik van scoringsrubrieken), en ondanks het feit dat studenten positief stonden tegenover bijvoorbeeld het gebruik van cijferweging, blijven ze weinig vertrouwen hebben in de beoordelingen van hun peers. Een summatieve peer-assessmentopzet die studenten als leerzaam én betrouwbaar ervaren, vereist dan ook verder onderzoek naar de randvoorwaarden waarbinnen dit kan geschieden. Het is bijvoorbeeld niet duidelijk in hoeverre de validiteit en betrouwbaarheid zouden toenemen of kunnen worden gecorrigeerd als de docent een aantal beoordelingen zelf voor zijn rekening zou nemen. Evenwel zou een hoge vastgestelde mate van overeenstemming tussen student- en docentbeoordelingen het vertrouwen van studenten in deze methode wel kunnen doen toenemen doordat de gepercipieerde validiteit hoger is (Segers, 2004). Cho, Schunn & Wilson (2006) maken daarbij de kanttekening dat een vanuit het docentperspectief objectief betrouwbaar en valide peer-assessmentproces niet noodzakelijkerwijze door studenten als betrouwbaar wordt ervaren. Ze pleiten voor meer onderzoek naar manieren waarop studentpercepties van de betrouwbaarheid en validiteit van peeroordelen kunnen worden verbeterd. Aangezien er geen direct causaal verband lijkt te zijn tussen feitelijke betrouwbaarheid en acceptatie door studenten, zijn twijfels over betrouwbaarheid dan wel validiteit waarschijnlijk geen goede reden voor docenten om peer assessment niet in te zetten.
Het vaststellen van een wegingsfactor voor individuele beoordelaars was erop gericht de 'niet-serieuze' beoordelaars minder mee te laten tellen voor het totale cijfer. In welke mate hierdoor een meer accurate beoordeling voor *alle* studenten tot stand is gekomen, is niet vastgesteld. Het blijft mogelijk dat door de wegingsprocedure een goede beoordelaar die extremere beoordelingsscores geeft, minder wordt meegeteld dan een groep beoordelaars die gematigder cijfers geven. Zoals in de paragraaf 'Cijferberekening en wegingsprocedure' is aangegeven, is ervoor gekozen de wegingsfactor voor beoordelaars (beperkt) mee te verrekenen in het cijfer van die beoordelaar. Dit kan tot gevolg hebben gehad dat studenten uit strategische overwegingen minder extreem zijn gaan beoordelen. Zelf vindt bijna de helft van de studenten dat hiervan enigszins sprake was. Dan gaat het wel om de studenten die de vragenlijst hebben ingevuld, en het feit dat de gemiddelde beoordelingsscores tussen week 3 en 5 vrijwel gelijk zijn gebleven, wijst niet op zo'n strategie. In combinatie met een afnemende spreiding over de weken van de cijfers binnen de beoordelingsgroepen (de beoordelaars werden het dus meer met elkaar eens over het cijfer dat een caseopdracht moet krijgen) zou dit er óók op kunnen duiden dat de beoordelaars over de weken beter getraind raken in het beoordelen c.q.

wellicht beter de scoringsrubrieken kunnen hanteren om tot een beoordeling te komen.
Het gebruik van een duidelijk en expliciet beoordelingsschema, in de vorm van een scoringsrubriek, bleek essentieel om een grote groep studenten zelfstandig cijfers te laten geven aan elkaar. Uit de ontwikkeling van de cijfers die studenten aan elkaar geven over de vijf weken, bleek ook dat de scores van week 2 hoger zijn dan die van week 3, 4 en 5. Dit zou verklaard kunnen worden doordat in week 3 scoringsrubrieken zijn ingevoerd, wat de studenten een meer uniforme basis kan hebben gegeven voor hun beoordelingen.
In deze studie is een effect vastgesteld van de studieachtergrond op de attitude tegenover peer assessment. Deze studentgroep bestond voor meer dan de helft uit voormalige hbo-studenten: zij zijn positiever dan bachelorstudenten over de inzet en leerwinst van peer assessment en hebben een groter vertrouwen in de accuraatheid van hun cijfer. Hoewel niet bekend is in hoeverre de hbo-studenten reeds specifieke ervaring hadden met peer assessment als werk- en beoordelingsvorm, komt hun hogere waardering mogelijk voort uit het feit dat ze uit een 'schoolsere' onderwijsomgeving komen en meer ervaring hebben met soortgelijke werkvormen. Segers (2004) wijst erop dat studenten zich de eerste keer niet erg gemakkelijk kunnen voelen met deze assessmentmethode.

De online ondersteuning van peer assessment
De derde onderzoeksvraag richtte zich op de mogelijkheden en beperkingen van de inzet van een online peer-assessmentinstrument bij grote groepen. Studenten stonden overwegend positief tegenover het gebruik van een online support-tool. Ook Wen & Tsai (2006) stelden eerder een positieve studentattitude vast tegenover de online ondersteuning van peer assessment.
Een softwaretoepassing is onmisbaar om peer assessment bij een grote studentgroep te introduceren. Zonder dit instrument zijn random toekenning van opdrachten aan (meerdere, eventueel anonieme) beoordelaars en management van het peer-assessmentproces praktisch niet haalbaar. Dit betekent niet noodzakelijkerwijze dat het online ondersteunen van peer assessment veel tijdbesparing oplevert voor de docent: het koppelen van studenten aan te beoordelen opdrachten, het ontwikkelen van een online feedbackprocedure en beoordelingscriteria, het oplossen van technische problemen en het bewaken van de kwaliteit van het beoordelingsproces zijn taken die doorgaans veel tijd in beslag nemen. Het is daarom belangrijk dat online peer-reviewsystemen ook zaken als kwaliteitsbewaking en cijfermanagement integreren. Zo is het bij de summatieve inzet van peer assessment wenselijk dat cijfers online gekalibreerd of gewogen kunnen worden (Hamer, Kell & Spence, 2007).
Ook andere vormen van kwaliteitsbewaking zijn vaak niet ingebouwd maar wel wenselijk, vooral bij die vormen van online peer assessment waar weinig ruimte is voor face-to-face bijeenkomsten waar de feedback verder toegelicht en besproken kan worden. Een mogelijkheid voor studenten om de kwaliteit van de ontvangen feedback online te beoordelen, kan de docent snel inzicht geven in waar zich problemen met (de kwaliteit van) beoordelaars voordoen en geeft de beoordelaar zelf ook beter inzicht in de kwaliteit van zijn beoordelingen. Het invoeren van een (online) revisieonderdeel in een peer-

assessmentopdracht, waarbij de ontvangen feedback dient te worden verwerkt alvorens het eindproduct wordt ingeleverd, kan ervoor zorgen dat het proces van feedback geven en ontvangen serieuzer wordt genomen dan in de hier beschreven opzet het geval was. Waar dit laatste ook gekoppeld wordt aan mogelijkheden tot (online) interactie tussen feedbackgever en -ontvanger, worden het proces en eindresultaat ook transparanter voor alle betrokkenen. Verder onderzoek dient dan uit te wijzen wanneer welke aspecten van peer assessment beter online dan wel face-to-face kunnen worden ondersteund.

Literatuur

Ballantyne, R., Hughes, K. & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment and Evaluation in Higher Education, 27*, 427-441.

Berg, B.A.M van den (2003). *Peer assessment in universitair onderwijs. Een onderzoek naar bruikbare ontwerpen*. Proefschrift Universiteit Utrecht.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In M. Segers, F. Dochy & E. Cascallar (red.), *Optimizing new modes of assessment: In search for qualities and standards* (pp. 13-37). Boston, Dordrecht en Londen: Kluwer.

Cho, K., Schunn, C.D. & Wilson, R.W. (2006). Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives. *Journal of Educational Psychology 98(4)*, 891-901.

Dochy, F., Admiraal, W. & Pilot, A. (2003). Peer- en co-assessment als instrument voor diepgaand leren: bevindingen en richtlijnen. *Tijdschrift voor Hoger Onderwijs, 21*, 220-229.

Dochy, F., Segers, M. & Sluijsmans, D.M.A. (1999). The use of self-, peer-, and co-assessment in higher education: a review. *Studies in Higher Education, 24*, 3, 331-350.

Dierick, S. & Dochy, F. (2001). New lines in edumetrics: New forms in assessment lead to new assessmentcriteria. *Studies in educational evaluation, 27*, 307-329.

Dierick, S., Dochy, F. & Watering, G. van de (2001). Assessment in het hoger onderwijs: over de implicaties van nieuwe toetsvormen voor de edumetrie. *Tijdschrift voor hoger onderwijs, 19*, 2-18.

Falchikov, N. & Goldfinch, J. (2000). Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research, 70*, 287-322.

Hamer, J., Kell, C. & Spence, F. (2007). Peer Assessment Using Aropä. In: *Proceedings Ninth Australasian Computing Education Conference (ACE2007)*, Ballarat, Australia. CRPIT, 66. Mann, S. & Simon (red.), ACS. 43-54.

Hamer, J., Ma, K.T.K. & Kwong, H.H.F. (2005). A Method of Automatic Grade Calibration in Peer Assessment. In: Proceedings Seventh Australasian Computing Education Conference (ACE2005), Newcastle, Australia. CRPIT, 42. Young, A. & Tolhurst, D. (red.), ACS. 67-72.

Nunnally, J.C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

Prins, F.J. et al. (2005). Formative peer assessment in a CSCL environment: A case study. *Assessment & Evaluation in Higher Education, 30*, 417-444.

Segers, M. (2004). Assessment en leren als een twee-eenheid: Onderzoek naar de impact van assessment op leren. *Tijdschrift voor Hoger Onderwijs, 22(4)*, 188-219.

Segers, M. & Dochy, F. (2001). New assessment forms in problem-based learning: The value added of the students' perspective. *Studies in Higher Education, 26*, 327-343.

Sluijsmans, D.M.A. (2002). *Student involvement in assessment: the training of peer assessment skills*. Heerlen: Open Universiteit Nederland.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249-276.

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Casacallar (red.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 55-87). Londen: Kluwer Academic Publishers.

Tseng, S. & Tsai, C. (2006). On-line peer assessment and the role of peer feedback: A study of a high school computer course. *Computers & Education, 49*, 1161-1174.

Wen, M.L. & Tsai, C. (2006). University students' perceptions of and attitude towards (online) peer assessment. *Higher Education, 51*, 27-44.

Appendix A: Tekst vragenlijst

1. Ik vind de software om de peer reviews mee uit te voeren gemakkelijk te gebruiken.
2. De tijd voor het maken van de cases was voldoende.
3. De tijd voor het beoordelen van de cases van anderen was voldoende.
4. De manier waarop mijn cijfer bij peer review tot stand komt, is mij duidelijk.
5. Ik vind peer review een goede methode om met de leerstof bezig te zijn.
6. Ik vind peer review een goede methode om mijn cijfer vast te stellen.
7. Ik ben het ermee eens dat beoordelaars verschillend gewogen worden in de beoordeling van mijn case.
8. Ik heb er vertrouwen in dat, doordat mijn cijfer voor een case verschillend wordt gewogen per beoordelaar, het juiste cijfer tot stand komt.
9. De criteria in het scoringsschema waarop ik het werk van medestudenten moest beoordelen, waren voldoende duidelijk.
10. Doordat er gewogen werd, ging ik meer gematigde cijfers geven aan mijn medestudenten.
11. Ik vond dat er over het algemeen veel variatie zat tussen de uitwerkingen van de cases van mijn medestudenten.
12. Voor het commentaar aan medestudenten heb ik argumenten uit het boek gehaald.
13. Ik was bij latere cases nog net zo gemotiveerd om medestudenten serieus te beoordelen als bij de eerste cases.
14. Ik ben er zeker van dat ik een juiste score heb toegekend aan mijn medestudenten.
15. Ik vond de scores die ik van mijn medestudenten kreeg over het algemeen adequaat.

16. Het commentaar van mijn medestudenten op mijn werk vond ik over het algemeen goed onderbouwd.
17. Ik heb het commentaar van medestudenten op mijn werk elke week nagelezen.
18. Ik heb veel geleerd van het vergelijken van mijn antwoorden op de cases met de antwoorden van mijn medestudenten.
19. Ik heb veel geleerd van het formuleren van commentaar op het werk van mijn medestudenten.
20. Ik heb veel geleerd van het lezen van commentaar van medestudenten op mijn eigen opdrachten.
21. Ik vond het motiverend om wekelijks een caseopdracht te maken.
22. Ik vond het leuk om door middel van het beoordelen van cases van medestudenten met de leerstof bezig te zijn.
23. Ik vind Consumer Behaviour een interessant vak.
24. De docent geeft op een boeiende manier les.